URCA Hugo Harari-Kermadec

2008-2009 harari@ecogest.ens-cachan.fr

Économétrie 2 : données qualitatives, probit et logit

I Un modèle pour données qualitatives

Cette section est fortement inspirée du cours de Christophe Hurlin.
On est confronté à des données qualitatives en micro-économie et en marketing, lorsque l'on étudie des choix (d'achat, de consommation, de comportement, de licenciement) ou des risques de défaillance (prêt). On peut prendre un exemple : pour une population d'étudiants en L3, on s'intéresse à l'événement "s'inscrire dans un master".

I.1 Le modèle dichotomique

Par modèle dichotomique, on entend un modèle statistique dans lequel la variable expliquée ne peut prendre que deux modalités (variable dichotomique). Il s'agit alors généralement d'expliquer la survenue ou non d'un évènement, ou d'un choix. Dans notre exemple, l'étudiant s'inscrit ou non en master.
On considère un échantillon de n individus d'indices $i = 1, .., n$. Pour chaque individu, on observe si un certain évènement s'est réalisé et l'on pose :

$$Y_i = \begin{cases} 1 & \text{si l'évènement s'est réalisé (l'étudiant s'inscrit)} \\ 0 & \text{si l'évènement ne s'est pas réalisé (pas d'inscription)} \end{cases}$$

On remarque ici le choix du codage (0, 1) qui est traditionnellement retenu pour les modèles dichotomique. En effet, celui-ci permet de définir la probabilité de survenue de l'évènement comme l'espérance de la variable Y, puisque :

$$E[Y_i] = \Pr(Y_i = 1) \times 1 + \Pr(Y_i = 0) \times 0 = \Pr(Y_i = 1).$$

L'espérance de Y_i donne donc la probabilité que l'étudiant s'inscrive en master.
L'objectif des modèles dichotomiques consiste alors à expliquer la survenue de l'événement considéré en fonction de K caractéristiques observées $(X_{i1}, \ldots, X_{iK})$ pour un individu i de l'échantillon, par exemple l'age de l'étudiant, son statut marital, s'il a des enfants, le niveau de vie des parents...

I.2 Un modèle linéaire ?

De manière générale, comme pour le modèle linéaire, on écrit pour les variables explicatives $X_i = (1, X_{i1}, \ldots, X_{iK})$ et pour les paramètres $\theta = (\theta_0, \theta_1, \ldots, \theta_K)'$, de sorte que $\theta_0 + \theta_1 X_{i1} + \cdots + \theta_K X_{iK} = X\theta$.
L'usage direct d'un modèle linéaire est voué à l'échec : écrire $Y_i = X_i\theta + \varepsilon$ impose à $X_i\theta + \varepsilon$ de ne prendre que les valeurs 0 et 1. Dans notre exemple, ça reviendrait à vouloir exprimer l'inscription en master comme une fonction linéaire de l'age et des autres variables explicatives.
Graphiquement, les valeurs de Y ne sont pas distribuées autour d'une droite, mais sur deux droites parallèles, $Y = 0$ et $Y = 1$.

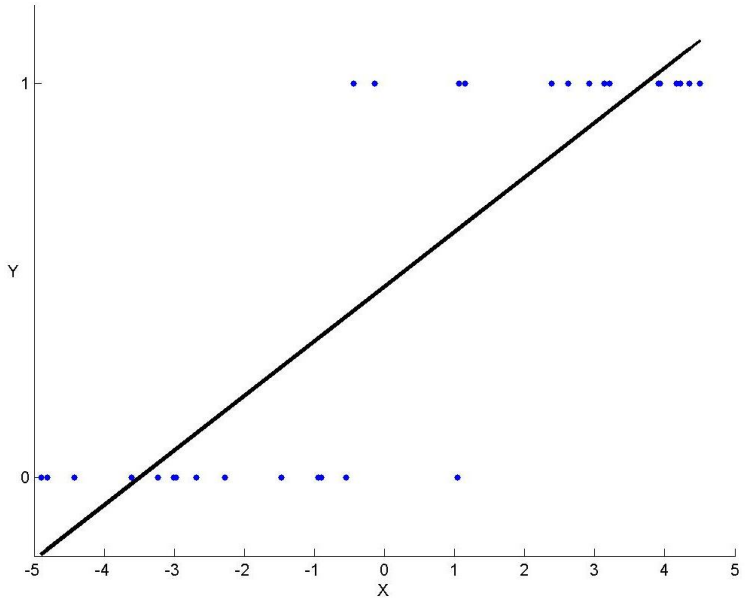

FIGURE 1 – Régression linéaire pour données qualitatives, avec $K = 1$.

En fait, par rapport au cadre d'usage du modèle linéaire, on observe beaucoup moins d'information. Ceci va apparaître grâce à l'introduction d'une "variable latente" Y^* :

$$Y_i = \begin{cases} 1 & Y_i^* \geq 0 \\ 0 & Y_i^* \leq 0. \end{cases} \qquad \text{où } Y_i^* = X_i\theta + \varepsilon_i,$$

c'est-à-dire

$$Y_i = \mathbb{1}_{X_i\theta+\varepsilon_i\geq 0}.$$

Pour utiliser les outils du modèle linéaire, il faudrait observer Y^*, ce qui n'est pas le cas. Il faut donc se résoudre à être moins ambitieux et à faire des hypothèses bien plus importantes.

I.3 Identification

Dans le cas gaussien, on va être amené à faire l'hypothèse très forte que les résidus sont **réduits** :

$$\varepsilon \sim \mathcal{N}(0, 1).$$

En effet, si l'on ne spécifie pas la variance de ε, on a un problème d'identification : les modèles

$$Y_i = \mathbb{1}_{0.2+3X_{i1}+\varepsilon_i\geq 0} \text{ avec } \varepsilon \sim \mathcal{N}(0, 1)$$
$$Y_i = \mathbb{1}_{0.4+6X_{i1}+\varepsilon_i\geq 0} \text{ avec } \varepsilon \sim \mathcal{N}(0, 4)$$

donnent exactement les mêmes observations. En supposant seulement que les résidus sont gaussiens, on est donc impossible d'estimer les paramètres θ_0 et θ_1.
On peut aussi choisir de spécifier que les résidus suivent la loi logistique, comme on va le voir au paragraphe suivant, l'important étant que la loi doit être totalement spécifiée.

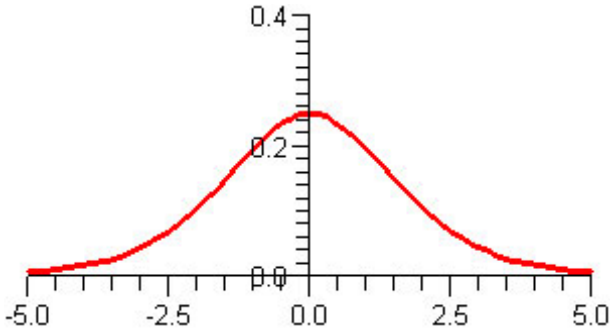

FIGURE 2 – La densité de la loi logistique.

I.4 Les modèles probit et logit

On cherche à expliquer les valeurs de Y grâce à X, c'est-à-dire à estimer la probabilité que $Y_i = 1$ sachant X_i (ou que $Y_i = 0$, ce qui revient au même). On remarque alors que :

$$\Pr(Y_i = 1|X_i) = \Pr(X_i\theta + \varepsilon_i \geq 0|X_i) = \Pr(X_i\theta \geq -\varepsilon_i|X_i) = F_{-\varepsilon}(X_i\theta).$$

La seule différence entre les modèles probit et logit est la spécification de F. Dans ces deux cas, la loi des résidus est symétrique, on peut donc remplacer $F_{-\varepsilon}$ par F_ε.

Probit Le modèle probit correspond à la spécification gaussienne introduite à la section précédente. F est donc la fonction de répartition d'un gaussienne centrée réduite, usuellement notée Φ :

$$F(X_i\theta) = \Phi(X_i\theta) = \int_{-\infty}^{X_i\theta} \frac{e^{-t^2/2}}{\sqrt{2\pi}} dt,$$

la densité correspondante, usuellement notée ϕ, est :

$$f(X_i\theta) = \phi(X_i\theta) = \frac{e^{-(X_i\theta)^2/2}}{\sqrt{2\pi}}.$$

Logit Le modèle Logit correspond à la loi logistique, introduite spécialement pour ce type de modèle, de fonction de répartition Λ :

$$F(X_i\theta) = \Lambda(X_i\theta) = \frac{e^{X_i\theta}}{1 + e^{X_i\theta}} = \frac{1}{1 + e^{-X_i\theta}},$$

la densité correspondante, usuellement notée λ, est :

$$f(X_i\theta) = \lambda(X_i\theta) = \frac{e^{-X_i\theta}}{(1 + e^{-X_i\theta})^2} = \Lambda(X_i\theta)(1 - \Lambda(X_i\theta)).$$

Il n'y a pratiquement pas de différence entre ces deux lois, l'introduction de la loi logistique étant simplement motivée par sa simplicité dans ce cadre.

I.5 Interprétation

Une fois le modèle estimé, on obtient des valeurs pour les paramètres (θ) qu'il faut interpréter. L'aspect essentiel est l'effet marginal de la j-ème variable X_{ij}, sur la probabilité de l'événement $Y = 1$ pour l'individu i. Cette effet s'écrit pour une variable X_{ij} continue (pour une variable explicative qualitative, il faut considérer un taux d'accroissement) :

$$\frac{\partial F(X_i\theta)}{\partial X_{ij}} = f(X_i\theta)\theta_j.$$

On a vu précédemment que les problèmes d'identification laisse peu de crédit à la valeur quantitative de θ_j, c'est donc surtout son signe que l'on va commenter. On peut donc "tirer" de ce modèle le signe de l'effet de X_{ij}.

Si $\theta_j > 0$, X_{ij} a un effet positif sur l'événement considéré.
Si $\theta_j < 0$, X_{ij} a un effet négatif sur l'événement considéré.

II Analyse statistique

II.1 Estimation par Maximum de Vraisemblance

On utilise la méthode du maximum de vraisemblance pour estimer nos paramètres. La vraisemblance s'écrit :

$$L(\theta) = \prod_{i=n}^{N} F(X_i\theta)^{Y_i}(1 - F(X_i\theta))^{1-Y_i}.$$

et donc la log-vraisemblance vaut :

$$\begin{aligned}\log L(\theta) &= \sum_{i=1}^{N} Y_i \log F(X_i\theta) + \sum_{i=1}^{N}(1 - Y_i)\log\left(1 - F(X_i\theta)\right) \\ &= \sum_{i:Y_i=1} \log F(X_i\theta) + \sum_{i:Y_i=0} \log\left(1 - F(X_i\theta)\right).\end{aligned}$$

Pour chaque modèle, on remplace F par sa valeur et l'on estime θ en résolvant la condition au premier ordre (la nullité du gradient de la log-vraisemblance). L'intérêt technique du modèle logit apparaît ici. En pratique, c'est bien sûr SAS qui s'occupe de cette étape.
Sous certaines conditions, l'estimateur du maximum de vraisemblance $\hat{\theta}$ est convergent et suit asymptotiquement une loi normale centrée sur la vraie valeur θ des paramètres et de matrice de variance covariance égale à l'inverse de la matrice d'information de Fisher $I(\theta)$ (l'opposée de l'espérance de la Hessienne de la log-vraisemblance). C'est la connaissance de cette loi asymptotique qui permet d'estimer les variances asymptotiques des estimateurs $\hat{\theta}_j$.

II.2 Tests

On peut obtenir des statistiques pivotales, c'est-à-dire des statistiques dont on connaît la loi asymptotique, qui permettent de tester des contraintes sur les coefficients, en particulier leur nullité. On obtient à chaque fois une statistique asymptotiquement χ^2, on compare donc les valeurs obtenues aux quantiles du χ^2.

Rapport de vraisemblance Dans le cadre de l'estimation par maximum de vraisemblance, le test le plus naturel consiste à construire un rapport de vraisemblance. Pour tester une contrainte de rang $p - r$ sur θ de dimension p, on utilise le résultat suivant :

$$LR = -2\left(\log L(\hat{\theta}) - \log L(\hat{\theta}^c)\right) \xrightarrow[N\to\infty]{\mathcal{L}} \chi^2_r,$$

où $\hat{\theta}^c$ est l'estimateur du maximum de vraisemblance sous la contrainte.

Score On peut aussi utiliser la nullité du score (aussi appelé test du multiplicateur de Lagrange), en mesurant la norme $||\cdot||_2$ du score évalué en $\hat{\theta}^c$:

$$\left.\frac{\partial \log L(\theta)}{\partial \theta}\right|'_{\theta=\hat{\theta}^c} I(\hat{\theta}^c)^{-1} \left.\frac{\partial \log L(\theta)}{\partial \theta}\right|_{\theta=\hat{\theta}^c} \xrightarrow[N\to\infty]{\mathcal{L}} \chi^2_r.$$

test de Wald Le test de Wald, proche du test de score, sert spécifiquement à tester la nullité d'un ou plusieurs coefficients, en particuliers de tous sauf la constante :

$$\frac{\hat{\theta}_j^2}{I_{jj}(\hat{\theta})} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_1 \text{ et } \sum_{k=1}^{K} \frac{\hat{\theta}_k^2}{I_{kk}(\hat{\theta})} \xrightarrow[n\to\infty]{\mathcal{L}} \chi^2_K.$$

II.3 Taux d'explication

On peut prendre du recul sur la modélisation et se demander simplement si notre modèle estimé est capable "d'expliquer" les observations. On se demande alors qu'elle valeur le modèle prédirait pour Y, sachant X. On calcule alors $F(X_i\hat{\theta})$ et on prédit $\hat{Y}_i = 1$ si la probabilité prédite pour l'occurrence de l'événement est supérieure à $1/2$, $\hat{Y}_i = 0$ sinon :

$$\hat{Y}_i = \mathbb{1}_{F(X_i\hat{\theta})>1/2}.$$

On peut alors calculer le taux de prédictions justes ($Y_i = \hat{Y}_i$).

III La procédure LOGISTIC

Les mots en majuscule sont des commandes SAS. Les mots en minuscule sont des noms donnés par l'utilisateur.
On suppose disposer d'un table "données" dans la librairie WORK (qui est la librairie par défaut).
"données" contient les variables Y, X1, X2, "sexe" et "poids". On suppose avoir ordonné la table par sexe.
On veut estimer le modèle expliquant Y par X1 et X2, pour chaque valeur de "sexe" séparément et avec les pondérations "poids".

$$\Pr(Y = 1|X1, X2) = F(\theta_0 + \theta_1 X1 + \theta_2 X2).$$

III.1 Entrée

PROC LOGISTIC DATA=données[1] ;
BY sexe [2] ;
MODEL Y = X1 X2 / LINK= LOGIT[3] ALPHA= 0.05[4] ;
OUTPUT OUT=sortie PROB=prédictions[5] XBETA=modélisés[6] ;
test1 : TEST INTERCEPT[7] + .5*X2 = 0 ;
test2 : TEST X1=X2 ;
WEIGHT poids ;
RUN ;

[1] Nom de la table a utiliser.
[2] BY sexe : lance la proc sur les sous populations définies par les valeurs de la variable sexe. Il faut avoir ordonner la table avant, avec :
PROC SORT DATA=données ; BY sexe ; run ;
[3] LINK= LOGIT (par défaut) ou PROBIT : l'inverse de F (Λ^{-1} ou Φ^{-1}).
[4] ALPHA= 0.05 (par défaut) ou 0.1 ou 0.01 : c'est le niveau de confiance des IC.
[5] Pour sauver dans une table "sortie" les données et les résultats, par exemple les valeurs de $\hat{Y}_i$.
[6] Calcule et sauve les valeurs de $X\hat{\theta}$. Sous SAS, le vecteur des paramètres est noté β, d'où le nom.
[7] INTERCEPT est la constante.

III.2 Sortie

On veut expliquer le ronflement par l'age, le sexe et la consommation régulière d'alcool. On tape donc :
Proc LOGISTIC DATA=donnees ;
MODEL ronfle=age sexe alcool/ LINK=PROBIT ;
RUN ;
On obtient :

```
                          Le Système SAS     10:54 Wednesday, January 17, 2007  31
                       The LOGISTIC Procedure
                     Informations sur le modèle

   Data Set                      WORK.DONNEES
   Response Variable             RONFLE               RONFLE
   Number of Response Levels     2
   Model                         binary probit
   Optimization Technique        Fisher's scoring

            Number of Observations Read         100
            Number of Observations Used         100

                      Profil de réponse
                 Valeur                  Fréquence
                ordonnée       RONFLE       totale
                       1            0           65
                       2            1           35

                Probability modeled is RONFLE=0.

                  État de convergence du modèle
        Convergence criterion (GCONV=1E-8) satisfied.

             Statistiques d'ajustement du modèle
             Coordonnée à l'origine              Coordonnée à l'origine
  Critère                 uniquement               et covariables
  AIC                        131.489                    122.209
  SC                         134.094                    132.629
  -2 Log L                   129.489                    114.209

         Test de l'hypothèse nulle globale : BETA=0
      Test                  Khi 2       DF     Pr > Khi 2
      Likelihood Ratio     15.2806       3         0.0016
      Score                13.7844       3         0.0032
      Wald                 13.6723       3         0.0034

       Analyse des estimations de la vraisemblance maximum
                                Erreur         Khi 2
Paramètre    DF    Estimation      std       de Wald    Pr > Khi 2
Intercept     1        2.7487   0.8057       11.6373        0.0006
AGE           1       -0.0385   0.0132        8.4432        0.0037
SEXE          1        0.1824   0.3740        0.2377        0.6258
ALCOOL        1       -0.1118   0.0468        5.7027        0.0169

  Association des probabilités prédites et des réponses observées
        Percent Concordant     72.2    Somers' D    0.445
        Percent Discordant     27.7    Gamma        0.446
        Percent Tied            0.1    Tau-a        0.205
        Pairs                2275    c            0.723
```

SAS commence par donner quelques informations descriptives sur les données et le modèle. On explique ci-dessous les principaux résultats.

Statistiques d'ajustement du modèle

Permet de comparer plusieurs modèles pour les mêmes données. Le meilleur modèle est celui pour lequel les critères sont les plus petits.

La première colonne donne les valeurs obtenues avec θ_0 uniquement, la seconde avec le modèle complet. On espère donc que les valeurs diminuent.

AIC
: Akaike Information Criterion : pénalisation de la log vraisemblance prenant en compte le nombre de variables explicatives.

SC
: Schwarz Criterion : pénalisation de la log vraisemblance prenant en compte le nombre de variables explicatives et le nombre de données.

$-2\log$ L
: $-2\log$ du maximum de la vraisemblance.

Test de l'hypothèse nulle globale : $\theta = 0$

Propose 3 tests pour la nullité de tous les coefficients. La première colonne la valeur de la statistique de test (asymptotiquement χ^2_p), la deuxième colonne rappelle p et la troisième donne la p-value, c'est-à-dire le quantile du χ^2 correspondant (on accepte la nullité si la p-value est grande).

Likelihood Ratio
: Le test basé sur le rapport de vraisemblance.

Score
: Le test basé sur le score.

Wald
: Le test de Wald.

Analyse des estimations de la vraisemblance maximum

Donne les valeurs estimées pour les $\hat{\theta}_j$ et des indications de leur significativité. La première ligne donne la constante $\hat{\theta}_0$. Les autres lignes donnent les coefficients de variables explicatives.

DF
: Rappelle la dimension du paramètre.

Estimation
: Donne $\hat{\theta}_j$.

Erreur std
: Donne l'estimation de l'écart-type : $\sqrt{I(\hat{\theta})_{jj}}$.

Khi 2 de Wald
: Donne la valeur du test de Wald pour la nullité du coefficient θ_j.

Pr > Khi 2
: Donne la p-value de ce test.

Association des probabilités prédites et des réponses observées

Donne des statistiques sur la justesse des prédictions $\hat{Y}_i$. Une paire est formée par deux individus ayant une réponse différente : $Y_{i_1} \neq Y_{i_2}$. Il y a concordance si les $\hat{Y}_i$ sont dans le même ordre que les Y_i. Si les ordres sont différents, la paire est discordante.

Percent Concordant
: Le pourcentage de paires concordantes.

Percent Discordant
: Le pourcentage de paires discordantes.

Percent Tied
: Le pourcentage de paires indéterminées.

Pairs
: Le nombre de paires.

Somers' D
: D de Somers : indice de la justesse des prédictions. Les 3 indices suivants jouent le même rôle. c'est indices sont compris entre 0 et 1, et l'on veut une mesure proche de 1.

Gamma
: Γ de Goodman-Kruskal.

Tau-a
: τ_a de Kendall.

c
: c de Hanley and McNeil 1982.

IV TP SAS

En tapant harari et reims dans google, vous devriez trouver ma page web vous concernant. Sinon, l'adresse est `http://www.crest.fr/ckfinder/userfiles/files/Pageperso/hharari/harari_fichiers/reims.htm`
On va s'attaquer au jeu de donnée télé-achat. L'objectif est d'expliquer si la vente est nulle ou non.

1. Mise en jambe
   (a) Importez les données et faites en une copie de sauvegarde.
   (b) Faire une analyse descriptive de la table.
2. Analyse statistique globale
   (a) Chercher le meilleur modèle possible pour expliquer le s'il y a eu vente ou non. Vous pouvez améliorer le modèle en choisissant au mieux les variables explicatives et la loi des résidus (gaussienne ou logistique).
   (b) Conjecturer un lien entre les coefficients de ce modèle et tester cette conjecture.
3. Analyse statistique différenciée
   (a) Reprendre l'étude séparément pour les jours de semaine et pour le week-end.
   (b) Trouve-t-on une différence avec l'analyse globale ?